A New Weighted NMF Algorithm For Missing Data Interpolation And Its Application To Speech Enhancement

Sushmita Thakallapalli, Suryakanth Gangashetty
Speech Processing Laboratory
IIIT, Hyderabad, India
{sushmita.t@research.iiit.ac.in},{svg@iiit.ac.in}

Nilesh Madhu
IDLab, Dept. Electronics & Information Systems
Ghent University - imec, Belgium
nilesh.madhu@ugent.be

***Abstract*—In this paper we present a novel weighted NMF (WNMF) algorithm for interpolating missing data. The proposed approach has a computational cost equivalent to that of standard NMF and, additionally, has the flexibility to control the degree of interpolation in the missing data regions. Existing WNMF methods do not offer this capability and, thereby, tend to overestimate the values in the masked regions. By constraining the estimates of the missing-data regions, the proposed approach allows for a better trade-off in the interpolation. We further demonstrate the applicability of WNMF and missing data estimation to the problem of speech enhancement. In this preliminary work, we consider the improvement obtainable by applying the proposed method to ideal binary mask-based gain functions. The instrumental quality metrics (PESQ and SNR) clearly indicate the added benefit of the missing data interpolation, compared to the output of the ideal binary mask. This preliminary work opens up novel possibilities not only in the field of speech enhancement but also, more generally, in the field of missing data interpolation using NMF.**

***Index Terms*—Weighted NMF, speech enhancement, binary mask, mask smoothing**

I. Introduction

The goal of traditional single-channel speech enhancement is to define a gain or mask function $\mathcal{G} \in \left[0, 1\right]$ in some chosen representation of the signal such that $\mathcal{G}$ is close to 1 in regions where the target speech is dominant and close to 0 (or some chosen threshold value) where the background noise is dominant. The chosen representation of the signal is generally some form of a time-frequency decomposition. Further the noise power spectrum is usually estimated from the noisy signal under the assumptions that this noise is uncorrelated with the target speech and has a stationarity span that is larger than that of the target speech. Several well known approaches to estimate the noise floor may be found in the literature [1]–[3]. The $\mathcal{G}$ can then be estimated by standard approaches in the literature, e.g. [4]–[7].

Another well-known method to obtain $\mathcal{G}$ is the *binary mask* which quantises the $\mathcal{G}$ to either 0 or 1, depending on the ratio between the speech and interference energy in a segment and a chosen decision threshold. The binary mask approach, coupled with oracle information on target and interference energy is often used to investigate the potential of single-channel speech enhancement approaches. Such an oracle-mask is termed the *ideal* binary mask (IBM) [8], [9] in the literature.

In all cases of single-channel speech enhancement, the $\mathcal{G}$ serves to highlight the 'reliable' regions of the noisy signal, i.e. regions that predominantly contain the target speech. However, irrespective of whether the $\mathcal{G}$ is obtained as an IBM or by any practical enhancement method, there are often regions of the target speech that are suppressed, either due to errors in the estimation of the noise floor (in practical approaches) or because the ratio of signal energy to interference energy fell below the set threshold for the IBM. This leads to 'holes' in the reconstructed signal, which can produce audible artefacts. The influence of these errors could be reduced by estimating such missing data points by some form of *interpolation*. We focus here, on the use of non-negative matrix factorisation (NMF) [10] as one such tool for missing data interpolation.

NMF allows a low-rank approximation of a large non-negative matrix by decomposing it as a product of two smaller nonnegative matrices. Such a decomposition gives interpretable representations that are used in various data analysis applications. When applied to a speech spectrogram, the NMF decomposition yields the *latent structure* and the *activations* of the basic frequency components. While NMF is a powerful data handling tool, it cannot be directly applied to matrices with *missing* data. Hence a variant of NMF called weighted NMF (WNMF), was developed to decompose data matrices with incomplete (missing) observations in the matrix. This approach treats an incomplete data matrix as being the product of an underlying full data matrix, multiplied with a binary mask where a '1' indicates an observation and '0' indicates missing data. WNMF was first proposed in [11], for the cost function based on the Euclidean distance measure and multiplicative update rules were derived for this case. However, for several signal matrices with a high dynamic range of the data (e.g. audio signals), the Kullback-Leibler (KL) divergence cost function is more appropriate [12]. Especially for audio, this cost function emphasises the perceptually important low energy, high frequency components.

However, extending the WNMF approach of [11] to the KL divergence measure leads to outliers/overestimation of the

component matrices of the NMF decomposition. To address this issue, an alternative approach [13] based on expectation maximisation (EM) was proposed. In this method, the parameters (basis and activation matrices) of the low-dimensional linear model are updated in the M-step and the missing entries are replaced by the NMF estimate of the data matrix in the E-step. Since missing data entries are directly replaced by the NMF estimate and the NMF decomposition is iterated on this 'filled-in' data matrix, two issues arise: (1) the extent of interpolation is not tunable, leading to overestimation of the missing values and (2) after each E-step, the EM-NMF requires a complete NMF (equivalent to hundreds of updates of the parameters) in the M-step, resulting in a high computational cost.

In this paper we propose a new WNMF approach based on the KL divergence, where the missing data regions are estimated in a constrained manner. By changing the weight on the constraint, the extent of interpolation in the missing-data regions can be controlled. Furthermore, the complexity of this method is equivalent to that of a conventional NMF approach based on multiplicative updates. Our goal in developing this method was to use NMF to interpolate among the poorly estimated regions of the denoised speech spectra, to further improve the clean speech spectral estimate after noise suppression. We therefore demonstrate the suitability of this method (as compared to the state-of-the-art EM-based WNMF) in an application of single-channel noise suppression by IBMs.

In the following we first describe the signal model and the state-of-the-art. We then propose our constrained WNMF approach and compare it with the state-of-the-art. Finally, we carry out an instrumental evaluation of the approach and demonstrate its potential for the case where the clean speech spectrum is estimated by an IBM. We note that this is preliminary work, and in the final section give the directions for future research.

II. NMF AND WEIGHTED NMF

Consider an $(M \times N)$ dimensional non-negative and real-valued data matrix $\mathbf{X}$. NMF seeks a factorisation of $\mathbf{X}$ into component non-negative matrices $\mathbf{W} \in \mathbb{R}^{M \times R}$ and $\mathbf{H} \in \mathbb{R}^{R \times N}$ such that $\mathbf{X} \approx \mathbf{WH}$. In such a representation, the matrices $\mathbf{W}$ and $\mathbf{H}$ are respectively termed the 'bases' and the 'activations'. The latent structure in $\mathbf{X}$ is assumed to be captured in $\mathbf{W}$. Further, if $R < (MN/(M+N))$, we obtain a low-rank approximation of $\mathbf{X}$. The cost functions commonly used to find the optimal $\mathbf{W}$ and $\mathbf{H}$ are based on the *Euclidean* distance or the *Kullback-Leibler* divergence. These are given in (1). Multiplicative update rules are typically used for the updates of the parameters.

$$\mathcal{J}^{\text{Euclidean}}(\mathbf{W},\mathbf{H}) = \frac{1}{2}\sum_{m,n}\Big(X_{mn} - \big(\mathbf{WH}\big)_{mn}\Big)^2 \quad (1)$$

$$\mathcal{J}^{\text{KL}}(\mathbf{W},\mathbf{H}) = \sum_{m,n}\Big(X_{mn}\log\left(\frac{X_{mn}}{\big(\mathbf{WH}\big)_{mn}}\right) - X_{mn} + \big(\mathbf{WH}\big)_{mn}\Big)$$

If, instead of $\mathbf{X}$ we had the matrix $\widetilde{\mathbf{X}}$ where not all elements mn of $\widetilde{\mathbf{X}}$ were 'observed', we have an *incomplete* representation of $\mathbf{X}$. Decomposing $\widetilde{\mathbf{X}} \approx \widetilde{\mathbf{W}}\widetilde{\mathbf{H}}$ would lead to the NMF trying to minimise the error across *all* entries, observed and unobserved, which leads to an underestimation of the parameters. To overcome this problem, a *weighted* NMF was proposed in [11] and was developed for the Euclidean distance measure as:

$$\mathcal{J}^{\text{WNMF}}(\mathbf{W},\mathbf{H}) = \frac{1}{2}\sum_{m,n}\mathcal{G}_{mn}\Big(\widetilde{X}_{mn} - \big(\mathbf{WH}\big)_{mn}\Big)^2, \quad (2)$$

where $\mathcal{G}_{mn}$ was a *binary weight* assigned to the missing-data matrix $\widetilde{\mathbf{X}}$, where a '1' signified that the corresponding data point is observed and '0' indicated that the data point was unobserved/missing. Multiplicative update rules were subsequently derived to estimate the parameters. However, this approach demonstrates poorer convergence and stability issues (especially when applied to the KL divergence measure). Further, in general, it tends to overestimate the data in the missing regions.

III. EM-WNMF

To deal with the drawbacks of the WNMF, a two-stage approach based on expectation maximisation (EM) was proposed in [13]. In this approach, the E-step corresponds to imputation where the missing data regions are filled-in using the current model estimate and the standard NMF multiplicative updates are applied on the filled-in matrix in the M-step. This two-stage approach is summarised below:

- E-Step: update the missing values using the previous NMF estimate.

$$\mathbf{Y} \leftarrow \boldsymbol{\mathcal{G}} \odot \widetilde{\mathbf{X}} + \big(\mathbf{1}^{M\times N} - \boldsymbol{\mathcal{G}}\big) \odot \mathbf{WH} \quad (3)$$

- M-Step: re-compute the NMF decomposition on the filled-in data matrix.

$$\mathbf{W} \leftarrow \mathbf{W} \odot \left(\frac{\frac{\mathbf{Y}}{\mathbf{WH}}\mathbf{H}^T}{\mathbf{1}\mathbf{H}^T}\right) \quad (4)$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \left(\frac{\mathbf{W}^T\frac{\mathbf{Y}}{\mathbf{WH}}}{\mathbf{W}^T\mathbf{1}}\right),$$

where $\mathbf{1}$ represents an $(M \times N)$ matrix of ones, $\odot$ represents the Hadamard product and the division is element-wise. This approach is carried out in multiple iterations of the E-and the M-steps. For each imputation step, the parameters $\mathbf{W}$ and $\mathbf{H}$ have to be re-estimated, requiring several hundred iterations of the conventional updates. This leads to a high computational cost, which can be somewhat improved by the approaches in [14]. Another issue with this approach is the following: since the missing/unknown values of the data matrix are replaced in their entirety by the low-rank model estimates, there is no means to control the degree of the interpolation, leading to over-estimation.

IV. Constrained WNMF (C-WNMF)

The above considerations lead us to propose the following constrained version of the KL-divergence based weighted NMF cost function:

$$\mathcal{J}^{\text{C-WNMF}}(\mathbf{W},\mathbf{H}) = \sum_{m,n} \mathcal{G}_{mn}\Big(\widetilde{X}_{mn}\log\left(\frac{X_{mn}}{(\mathbf{WH})_{mn}}\right) - \widetilde{X}_{mn} + (\mathbf{WH})_{mn}\Big) + \lambda(1-\mathcal{G}_{mn})(\mathbf{WH})_{mn}\,, \tag{5}$$

where λ is a parameter that allows us to control the degree of interpolation in the missing-data regions. From this cost function, the following multiplicative update rules can be derived:

$$\mathbf{W} \leftarrow \mathbf{W} \odot \left(\frac{\dfrac{\boldsymbol{\mathcal{G}} \odot \widetilde{\mathbf{X}}}{\mathbf{WH}}\mathbf{H}^T}{\Big(\boldsymbol{\mathcal{G}} + \lambda\big(\mathbf{1}-\boldsymbol{\mathcal{G}}\big)\Big)\mathbf{H}^T}\right) \tag{6}$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \left(\frac{\mathbf{W}^T\dfrac{\boldsymbol{\mathcal{G}} \odot \widetilde{\mathbf{X}}}{\mathbf{WH}}}{\mathbf{W}^T\Big(\boldsymbol{\mathcal{G}} + \lambda\big(\mathbf{1}-\boldsymbol{\mathcal{G}}\big)\Big)}\right),$$

where the various operators have the same meaning previously ascribed.

It is easy to see that when there is no missing data, the above rules converge to the traditional (non-weighted) NMF updates.

V. Application to speech enhancement

First, consider a speech signal (magnitude) spectrum $\mathbf{S}$ where parts of the spectrum are 'missing' (i.e. by application of a random binary mask). The estimation of this spectrum by the various NMF methods is depicted in the Figure 1, where it can be seen that in addition to the lower computational cost (equivalent to that of a standard NMF), the C-WNMF approach also has a better interpolation capability. In addition, the parameter λ can be chosen to trade-off the amount of interpolation.

This synthetic example sets the context for the following realistic application to speech enhancement. Consider a speech signal $s(n)$ corrupted by additive background noise $v(n)$. The noisy mixture is then denoted as $x(n) = s(n) + v(n)$. Speech enhancement and noise suppression is usually carried out in the short-time Fourier representation of the signal, obtained by computing the discrete Fourier transform (DFT) on windowed and overlapped segments of the signal. This leads to the following representation:

$$X(\ell,k) = S(\ell,k) + V(\ell,k)\,, \tag{7}$$

where ℓ and k represent the time-frame and frequency bin indices respectively. Enhancement is carried out by weighting $X(\ell,k)$ by the gain function $\mathcal{G}\big(\ell,k\big) \in [0,1]$, such that $\mathcal{G}$ is high in speech-dominated time-frequency regions and low in noise-dominated regions. As a first application to demonstrate the potential of WNMF for spectral interpolation, we choose the *ideal* binary mask (IBM) as the enhancement function. The IBM gain function is based on oracle knowledge and is defined as in (8).

$$\mathcal{G}^{\text{IBM}}(\ell,k) = \begin{cases} 1 & |S(\ell,k)|^2 > \Gamma|V(\ell,k)|^2 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

with Γ being a threshold parameter.

Applying the IBM to the noisy amplitude spectrum $|X(\ell,k)|$ yields the 'missing-data' matrix we use for the weighted NMF. We consider $\widetilde{\mathbf{X}}$ to be the IBM-masked spectrum:

$$\widetilde{X}(\ell,k) = \mathcal{G}(\ell,k)|X(\ell,k)|. \tag{9}$$

The WNMF decomposition of this yields $\widetilde{\mathbf{X}} \approx \mathbf{WH}$, from which we then generate a new gain function by:

$$\mathcal{G}^{\text{WNMF}}(\ell,k) = \min\left(1, \frac{(\mathbf{WH})_{\ell,k}}{|X(\ell,k)|}\right) \tag{10}$$

Since this mask should only estimate the missing regions of the spectrum, the final gain function is obtained as:

$$\mathcal{G}^{\text{Interp}}(\ell,k) = \big(1-\mathcal{G}^{\text{IBM}}(\ell,k)\big)\mathcal{G}^{\text{WNMF}}(\ell,k) + \mathcal{G}^{\text{IBM}}(\ell,k) \tag{11}$$

VI. Experiments and results

A. Experimental set-up

To demonstrate the benefit of WNMF-based interpolation, we compare the performance of the IBM against the EM-WNMF and the C-WNMF based interpolation applied to (10) and (11). For this we consider input speech files consisting of 4 male and 4 female voices taken from the TSP-speech database [15]. Pink and babble noise from the ETSI noise database [16] are added to these clean speech signals at SNRs of -5 and $0\,$dB.

We conduct our experiments at a sample rate $f_s = 16\,$kHz, a DFT size of $K = 512$ samples, and a frame shift of 25%. A periodic square root Hann window is employed for both analysis and overlap-add synthesis. A dictionary size of 40 is chosen for the NMF approaches. C-WNMF is further tested with two constraint parameters $\lambda \in \{0.1, 0.4\}$.

The instrumental measures chosen are PESQ and segmental SNR improvement. The segmental SNR improvement (SegSNRi) is computed as:

$$\text{SegSNRi} = \text{SegSNR}_{\text{out}} - \text{SegSNR}_{\text{in}} \tag{12}$$

where $\text{SegSNR}_{\text{in}}$ is the average input segmental SNR, measured using the clean speech and the noise scaled to the input SNR value. $\text{SegSNR}_{\text{out}}$ is the average output segmental SNR measured by applying the resulting gain function *separately* to the clean speech and the scaled noise signal. This measure gives an indication of the amount of noise suppression. The SegSNR is defined according to [17].

Good speech component quality is reflected in a high PESQ mean opinion score (MOS-LQO) [18] which is applied to the *filtered* clean speech component, with the clean speech component as a reference. We do not measure PESQ on

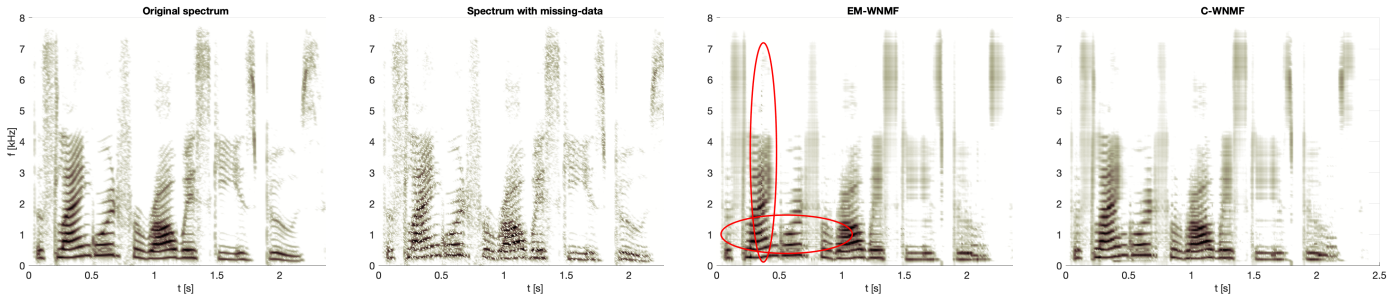

Fig. 1. Toy example illustrating the application of WNMF for missing data interpolation in speech enhancement. The first plot indicates the clean (underlying) spectrum. The second plot shows the spectrum with missing components (generated by a random binary mask), the third and fourth plots depict the low-rank signal reconstruction by EM-WNMF and the proposed C-WNMF. The artefacts generated in the EM-WNMF due to over-estimation are highlighted for convenience. Note, also, the better signal estimate for the C-WNMF, clearly visible for the fricatives around $t \approx 1.8\,\mathrm{s}$ and $t \approx 2.3\,\mathrm{s}$. Thus, in addition to lower computational complexity, the C-WNMF also demonstrates a better interpolation capability. A constraint of $\lambda = 0.1$ was chosen for C-WNMF.

the enhanced signal since PESQ has not been validated for artefacts caused by noise reduction techniques. This is in line with the reasoning in [19]

B. Results and Observations

The table below shows the PESQ and SegSNRi scores obtained after noise suppression using IBM and the interpolated masks using the EM-WNMF and the proposed C-WNMF approaches.

TABLE I
AVERAGE PESQ AND SEGSNRI SCORES OBTAINED AFTER NOISE SUPPRESSION USING THE IBM, THE EM-WNMF APPROACH AND THE PROPOSED C-WNMF METHOD FOR GAIN INTERPOLATION FOR THE 0DB AND -5DB CASES. THE DICTIONARY SIZE IS 40.

	IBM	EM-WNMF	C-WNMF ($\lambda = 0.1$)	C-WNMF ($\lambda = 0.4$)
SNR = 0dB				
PESQ	1.83	2.81	2.22	2.05
SegSNRi	7.60	4.43	6.32	6.93
SNR = -5dB				
PESQ	1.82	2.70	2.16	2.00
SegSNRi	8.05	5.20	7.15	7.50

From Table I we observe that the PESQ scores of the proposed C-WNMF are higher than those of IBM. Since this is PESQ measured on the filtered clean-speech signal it indicates that the C-WNMF preserves more speech components. At the same time, the SegSNRi of C-WNMF is comparable to that of the IBM approach, indicating that the improvement in the speech quality is not at the cost of noise suppression. In comparison, the EM-WNMF method shows a much higher PESQ score, but also correspondingly much lower SegSNRi. This indicates an over-estimation of the missing regions, leading to noise-vocoded output. When comparing the C-WNMF results for different values of the λ, we see how this provides a control over the interpolation. Increasing the λ leads to less interpolation (evidenced by the drop in the PESQ scores and an increase in the SegSNRi scores). Thus, we come closer to the performance of the IBM as the λ increases. Note, however, that there is still some measure of interpolation being done, hence even for $\lambda = 1$, we expect the performance of interpolated gain function using C-WNMF to be better than that of the IBM in terms of speech quality, while the noise suppression would be comparable to that of the IBM.

VII. CONCLUSIONS

We have introduced a new weighted NMF (WNMF) approach for missing-data interpolation, which allows control on the level of interpolation of the missing data. This approach, termed constrained WNMF (C-WNMF), has been demonstrated to be superior to the state-of-the-art, both in terms of computational cost as well as interpolation capability. We have further demonstrated its use in a speech enhancement framework, considering an oracle-knowledge based gain function (the ideal binary mask (IBM)). Note that we chose the IBM for this analysis since it allows us to analyse our method by disregarding (for the present) possible interactions with other components (e.g., noise power estimation, gain computation) in a real noise-suppression framework. We note that this paper serves primarily to demonstrate the benefit of the new C-WNMF method compared to the state-of-the-art and is a first study on the use of NMF in this context of speech enhancement. There are several opportunities to expand upon this topic and in the future, we would like to investigate and extend this method for the case of soft-masks and evaluate it for the case where the masks are not ideal but estimated within a single-channel noise suppression framework.

REFERENCES

[1] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, July 2001.

[2] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication*, vol. 48, pp. 220–231, 2006.

[3] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *Proceedings of the IEEE International Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2011, pp. 145–148.

[4] J. Porter and S. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1984.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[6] P. J. Wolfe and S. J. Godsill, "Simple alternatives to the ephraim and malah suppression rule for speech enhancement," in *IEEE Signal processing workshop on statistical signal processing*, 2001.
[7] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2098–2108, Nov. 2006.
[8] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *in Speech Separation by Humans and Machines*. Kluwer, 2005, pp. 181–197.
[9] ——, "Time–frequency masking for speech separation and its potential for hearing aid design," *Trends in amplification*, pp. 332–353, Oct. 2008.
[10] D. D. Lee and S. H. Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of the 13th International Conference on Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2000, pp. 535–541.
[11] Y. Mao and L. K. Saul, "Modeling distances in large-scale networks by matrix factorization," in *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*, ser. IMC '04, 2004, pp. 278–287.
[12] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional models for audio processing: Uncovering the structure of sound mixtures," *IEEE Signal Processing Magazine*, vol. 32, pp. 125–144, March 2015.
[13] S. Zhang, W. Wang, J. Ford, and F. Makedon, "Learning from incomplete ratings using non-negative matrix factorization," in *Proceedings of the 2006 SIAM International Conference on Data Mining*, 2006, pp. 549–553.
[14] Y. Kim and S. Choi, "Weighted nonnegative matrix factorization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2009, pp. 1541–1544.
[15] P. Kabal, "TSP speech database," Telecommunications and Signal Processing Laboratory, McGill University, Canada, Tech. Rep., 2002. [Online]. Available: http://www-mmsp.ece.mcgill.ca/Documents/Data/
[16] ETSI, "Eg 202 396-1: Speech processing, transmission and quality aspects (stq); speech quality performance in the presence of background noise; part 1: Background noise simulation technique and background noise database, european telecommunications standards institute," Sep. 2008. [Online]. Available: http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html
[17] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model," *EURASIP Journal on Advances in Signal Processing*, no. 7, pp. 1110 – 1126, May 2005.
[18] ITU, "Rec. p.862: Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, international telecommunication union, telecommunication standardization sector (itu-t)," Feb. 2001.
[19] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "Two-stage speech enhancement with manipulation of the cepstral excitation," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, March 2017, pp. 106–110.